

Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels

P. Alquier^{*}, N. Friel^{*1}, R. G. Everitt[†], A. Boland^{*}.

^{*}*School of Mathematical Sciences and Insight: The National Centre for Big Data Analytics, University College Dublin, Ireland.*

[†]*Department of Mathematics and Statistics, University of Reading, UK.*

March 25, 2014

Abstract

Monte Carlo algorithms often aim to draw from a distribution π by simulating a Markov chain with transition kernel P such that π is invariant under P . However, there are many situations for which it is impractical or impossible to draw from the transition kernel P . For instance, this is the case with massive datasets, where it is prohibitively expensive to calculate the likelihood and is also the case for intractable likelihood models arising from, for example, Gibbs random fields, such as those found in spatial statistics and network analysis. A natural approach in these cases is to replace P by an approximation \hat{P} . Using theory from the stability of Markov chains we explore a variety of situations where it is possible to quantify how 'close' the chain given by the transition kernel \hat{P} is to the chain given by P . We apply these results to several examples from spatial statistics and network analysis.

Keywords and Phrases: Markov chain Monte Carlo; Pseudo-marginal Monte Carlo; intractable likelihoods.

1 Introduction

There is considerable interest in the analysis of statistical models with difficult to evaluate or intractable likelihood functions. Such models occur in a diverse range of contexts including spatial statistics, social network analysis, statistical genetics, finance and so on. The challenges posed by this class of models has led to the development of important theoretical and methodological advances in statistics. For example, Geman and Geman (1984) developed the Gibbs sampler to sample from an Ising model for application in image analysis. More recently, the area of approximate Bayesian computation has emerged to deal with situations where the likelihood is not available for evaluation, but where it is possible to simulate from the likelihood function. This area has generated much activity in the literature. See (Marin *et al* 2012) for a recent survey.

In many applications in statistics, well known theoretically efficient estimators are not available in practice for computational reasons. For example:

¹Address for correspondence: nial.friel@ucd.ie

1. large datasets: the sample size ℓ is too large. This situation is very common nowadays as huge databases can be stored at no cost. For example: in genomics the cost of sequencing has fallen by a factor of 10^5 in past decade and a half. This has led to the wide availability of sequence data - the recently announced Personal Genome Project UK aims to sequence 10^5 human genomes, each consisting of 3×10^8 bases.
2. high-dimensional parameter spaces: the sample size ℓ might be reasonable, but the number of variables p is too large. For example: data assimilation in numerical weather prediction, in which the size of the state space is typically 10^9 .
3. intractable models: the likelihood / regression / classification function is not available in closed form and each evaluation is computationally demanding. Common examples are: in the statistical modelling of large numbers of linked objects, leading to the intractable likelihood in graphical models, which is the main focus of the applications in this paper.

A new point of view in statistics emerged to address these challenging situations: to focus on the computational aspects first, by proposing a fast enough algorithm to deal with the data. In some way, this mean that we replace the traditional definition of an *estimator* as a *measurable function of the data* by an *algorithm able to proceed with the data*. However, this does not mean that we should forget the theoretical properties of this estimator: a study of its properties is necessary. A typical example is Tibshirani’s LASSO estimator (Tibshirani 1996), it became successful as the first estimator available in linear regression when p is very large ($> 10^6$), only later, were conditions provided to ensure its theoretical optimality. See (Bühlmann and Van de Geer 2011) for a survey. This idea to consider the algorithm as the definition of an estimator is pushed further in (Valiant 1984; Bottou and Bousquet 2011) among others.

This situation also appears in Bayesian statistics; while some Bayesian estimators can be efficiently approximated by MCMC methods such as the Metropolis-Hastings algorithm, sometimes, this is not possible because the acceptance ratio in the algorithm cannot be evaluated – indeed this is the focus of our paper. It is intuitive to replace this ratio by an estimate or an approximation. Nicholls *et al.* (2012), Andrieu and Roberts (2009) and Liang and Jin (2011) considered this idea for models with intractable likelihood. Korattikara *et al.* Both Bardenet *et al.* (2014) and Korattikara *et al.* (2014) applied this idea in the case where the sample size ℓ is too large to prohibit many evaluations of the likelihood. One might also view situations in which an approximating model is used (such as approximate Bayesian computation) as a special case of this general view, although such examples are not considered in this paper.

In this paper, we propose a general approach to “noisy” or “inexact” MCMC algorithms. In Section 2, we describe the main idea and provide a result, due to Mitrophanov, that gives a theoretical justification of the algorithm in many situations. Our results gives bounds on the distance, with respect to the total variation norm, between an “ideal” chain which leaves the target distribution invariant and a noisy chain which approximates the target distribution. We then study the special cases of a noisy version of the Exchange algorithm (Murray *et*

al. (2006)), and discretized Langevin Monte Carlo in Section 3. For these noisy algorithms we prove that the total variation distance decreases with the number of iterations, N , of the randomisation step in the noisy algorithm, and find a bound on this distance in terms of N . We study in detail an application to intractable likelihood problems in Section 4.

2 Noisy MCMC algorithms

In many practical situations, useful statistical estimators can be written as

$$\hat{\theta} = \int_{\Theta} \theta \pi(d\theta)$$

for some probability distribution π . This is for example the case in Bayesian statistics where π is the posterior distribution of θ given the data, but estimators under this form appear in other situations, e.g. the exponentially weighted aggregate (Dalalyan and Tsybakov 2012). More generally, one might want to estimate functionals of the form

$$\int_{\Theta} f(\theta) \pi(d\theta)$$

for some function f . A very popular approach in this case is the family of MCMC algorithms. The idea is simulate a Markov Chain $(\theta_n)_{n \in \mathbb{N}}$ with transition kernel P such that π is invariant under P : $\pi P = \pi$. We then use the approximation

$$\frac{1}{N} \sum_{n=1}^N f(\theta_n) \simeq \int_{\Theta} f(\theta) \pi(d\theta). \quad (1)$$

Of course, in order for such an approximation to be useful, we need more than the requirement that $\pi P = \pi$. A very useful property in this respect is so-called exponential ergodicity for which it holds that

$$\sup_{\theta_0} \|\delta_{\theta_0} P^n - \pi\| \leq C \rho^n,$$

for some $C < \infty$ and $\rho < 1$, where $\|\cdot\|$ is the total variation distance. Meyn and Tweedie (1993) detail conditions on P to ensure exponential ergodicity, and show theoretical results that ensure that (1) holds, in some sense.

However, there are many situations where there is a natural kernel P such that $\pi P = \pi$, but for which it is not computationally feasible to draw $\theta_{n+1} \sim P(\theta_n, \cdot)$ for a fixed θ_n . For these cases a natural approach is to replace P by an approximation \hat{P} so that when the approximation is good we hope that \hat{P} is “close” to P in some sense. Of course, in general we will have $\pi \hat{P} \neq \pi$, but we will show that it is nevertheless useful to ask the question whether it is possible to produce a Markov chain with an upper bound $\|\delta_{\theta_0} \hat{P}^n - \pi\|$?

It turns out that a useful answer to this question is given by the study of the stability of Markov chains. There have been a long history of research on this topic, we refer the reader

to the monograph by Kartashov (1996) and the references therein. Here, we will focus on a more recent method due to Mitrophanov (2005). In order to measure the distance between P and \hat{P} recall the definition of the total variation measure between two kernels:

$$\|P - \hat{P}\| := \sup_{\theta \in \Theta} \|\delta_\theta P - \delta_\theta \hat{P}\|.$$

Theorem 2.1 (Corollary 3.1 page 1006 in (Mitrophanov 2005)) *Let us assume that*

- **(H1)** *the Markov chain with transition kernel P is uniformly ergodic:*

$$\sup_{\theta_0} \|\delta_{\theta_0} P^n - \pi\| \leq C \rho^n$$

for some $C < \infty$ and $\rho < 1$.

Then we have, for any $n \in \mathbb{N}$, for any starting point θ_0 ,

$$\|\delta_{\theta_0} P^n - \delta_{\theta_0} \hat{P}^n\| \leq \left(\lambda + \frac{C \rho^\lambda}{1 - \rho} \right) \|P - \hat{P}\|$$

where $\lambda = \left\lceil \frac{\log(1/C)}{\log(\rho)} \right\rceil$.

This result serves as the basis for our paper. Practically, it says that the total variation distance between two Markov chains each of which have the same initial state, θ_0 , is less than or equal to a constant times the total variation distance between the kernels P and \hat{P} . It is interesting that this bound is independent of the number of steps n of the Markov chain.

The main purpose of this article is to show that there are many useful situations where this result can provide approximate strategies with the guarantee of theoretic convergence to the target distribution.

2.1 Noisy Metropolis-Hastings

The Metropolis-Hastings (M-H) algorithm, sequentially draws candidate observations from a distribution, conditional only upon the last observation, thus inducing a Markov chain. The M-H algorithm is based upon the observation that a Markov chain with transition density $P(\theta, \phi)$ and exhibiting detailed balance for π ,

$$\pi(\theta|\mathbf{y})P(\theta, \phi) = \pi(\phi|\mathbf{y})P(\phi, \theta),$$

has stationary density, $\pi(\theta)$.

In some applications, it is not possible to compute the ratio $\alpha(\theta'|\theta)$. In this case it seems reasonable to replace the ratio with an approximation or an estimator. For example, one could draw $y' \sim F_{\theta'}(\cdot)$ for some suitable probability distribution $F_{\theta'}(\cdot)$ and estimate the ratio α by $\hat{\alpha}(\theta'|\theta, y')$. This gives the ‘noisy’ Metropolis-Hastings algorithm in algorithm 2.

Algorithm 1 Metropolis-Hastings algorithm

for $n = 0$ to I **do**
 Draw $\theta' \sim h(\cdot|\theta_n)$

 Set $\theta_{n+1} = \theta'$ with probability $\min(1, \alpha(\theta', \theta_n))$

 where $\alpha(\theta', \theta_n) = \frac{\pi(\theta'|y)h(\theta_n|\theta')}{\pi(\theta_n|y)h(\theta'|\theta_n)}$

 Otherwise, set $\theta_{n+1} = \theta_n$.
end for

Algorithm 2 Noisy Metropolis-Hastings algorithm

for $n = 0$ to I **do**
 Draw $\theta' \sim h(\cdot|\theta_n)$

 Draw $y' \sim F_{\theta'}(\cdot)$

 Set $\theta_{n+1} = \theta'$ with probability $\min(1, \hat{\alpha}(\theta', \theta_n, y'))$

 Otherwise, set $\theta_{n+1} = \theta_n$.
end for

Note that $\hat{\alpha}(\theta', \theta, y')$ can be thought of as a randomised version of $\alpha(\theta', \theta)$ and as we shall see from the convergence result below, in order for this to yield a useful approximation, we require that $|\hat{\alpha}(\theta', \theta, y') - \alpha(\theta', \theta)|$ is small. Here we let \hat{P} denote the transition kernel of the Markov Chain resulting from Algorithm 2. Of course there is no reason for π to be invariant under \hat{P} , however we show under certain conditions that using an approximate kernel will yield a Markov chain which will approximate the true density. Moreover, we provide a bound on the distance between the Markov chain which targets π and the Markov chain resulting from \hat{P} .

2.1.1 Theoretical guarantees for Noisy Metropolis-Hastings

We now provide an application of Theorem 2.1 to the case of an approximation to the true transition kernel arising from Algorithm 2.

Corollary 2.2 *Let us assume that*

- **(H1)** *the Markov chain with transition kernel P is uniformly ergodic holds,*
- **(H2)** $\hat{\alpha}(\theta|\theta', y')$ *satisfies:*

$$\mathbb{E}_{y' \sim F_{\theta'}} |\hat{\alpha}(\theta, \theta', y') - \alpha(\theta, \theta')| \leq \delta(\theta, \theta'). \quad (2)$$

Then we have, for any $n \in \mathbb{N}$, for any starting point θ_0 ,

$$\|\delta_{\theta_0} P^n - \delta_{\theta_0} \hat{P}^n\| \leq \left(\lambda + \frac{C\rho^\lambda}{1-\rho} \right) \sup_{\theta} \int d\theta' h(\theta'|\theta) \delta(\theta, \theta'),$$

where $\lambda = \left\lceil \frac{\log(1/C)}{\log(\rho)} \right\rceil$.

All the proofs are given in Section A. The proof of Corollary 2.2 relies on the result by Mitrophanov (2005). Note, for example, that when the upper bound (2) is uniform, ie $\delta(\theta, \theta') \leq \delta < \infty$, then we have that

$$\|\delta_{\theta_0} P^n - \delta_{\theta_0} \hat{P}^n\| \leq \delta \left(\lambda + \frac{C\rho^\lambda}{1-\rho} \right).$$

Obviously, we expect that $\hat{\alpha}$ is chosen in such a way that $\delta \ll 1$ and so in this case, $\|\delta_{\theta_0} P^n - \delta_{\theta_0} \hat{P}^n\| \ll 1$ as a consequence. In which case, letting $n \rightarrow \infty$ yields

$$\limsup_{n \rightarrow \infty} \|\pi - \delta_{\theta_0} \hat{P}^n\| \leq \delta \left(\lambda + \frac{C\rho^\lambda}{1-\rho} \right).$$

Remark 2.1 *Andrieu and Roberts (2009) derived a special case of this result for a given approximation of the acceptance ratio α using their pseudo-marginal approach. We explore this more in section 2.4.*

Remark 2.2 *Another approach, due to Nicholls et al. (2012), gives a lower bound on the first time such that the chain produced by the Metropolis-Hastings algorithm and its noisy version differ, based on a coupled Markov Chains argument.*

Remark 2.3 *Note that a deterministic version of this result also holds in situations where one could replace $\alpha(\theta', \theta)$ by a deterministic approximation $\hat{\alpha}(\theta', \theta)$.*

We will show in the examples that follow in Section 3 that, when $\hat{\alpha}$ is well chosen, it can be quite easy to check that Hypothesis **(H2)** holds. On the other hand, it is typically challenging to check that Hypothesis **(H1)** holds. A nice study of conditions for geometric ergodicity of P is provided by Meyn and Tweedie (1993) and Roberts and Tweedie (1996b).

2.2 Noisy Langevin Monte Carlo

The Metropolis-Hastings algorithm can be slow to explore the posterior density, if the chain proposes small steps it will require a large number of moves to explore the full density, conversely if the chain proposes large steps there is a higher chance of moves being rejected so it will take a large amount of proposed moves to explore the density fully. An alternative Monte Carlo method is to use Stochastic Langevin Monte Carlo (Welling and Teh 2011). The Langevin diffusion is defined by the stochastic differential equation (SDE)

$$d\theta(t) = \nabla \log \pi(\theta(t)) dt/2 + db(t),$$

Algorithm 3 Langevin algorithm

```

for  $n = 0$  to  $I$  do
  Set  $\theta_{n+1} = \theta_n + \frac{\Sigma}{2} \nabla \log \pi(\theta_n) + \eta$  ,  $\eta \sim N(0, \Sigma)$ ,
end for

```

where $db(T)$ denotes a D-dimensional Brownian motion. In general, it is not possible to solve such an SDE, and often a first order Euler discretization of the SDE is used to give the discrete time approximation

However convergence of the sequence $\{\theta_n\}$ to the invariant distribution is not guaranteed for a finite step size Σ due to the first-order integration error that is introduced. It is clear that the Langevin algorithm produces a Markov chain and we let P_Σ denote the corresponding transition kernel. Note that, we generally don't have $\pi(\cdot|y)P_\Sigma = \pi(\cdot|y)$ nor $\delta_{\theta_0}P_\Sigma \rightarrow \pi(\cdot|y)$, however, under some assumptions, $\delta_{\theta_0}P_\Sigma \rightarrow \pi_\Sigma$ for some π_Σ close to π when Σ is small enough, we discuss this in more detail below.

In practice, it is often the case that $\nabla \log \pi(\theta_n)$ cannot be computed. Here again, a natural idea is to replace $\nabla \log \pi(\theta_n)$ by an approximation or an estimate $\hat{\nabla}^{y'} \log \pi(\theta_n)$, possibly using a randomization step $y' \sim F_{\theta_n}$. This yields what we term a noisy Langevin algorithm.

Algorithm 4 Noisy Langevin algorithm

```

for  $n = 0$  to  $I$  do
  Draw  $y_{\theta_n} \sim F_{\theta_n}(\cdot)$ .

  Set  $\theta_{n+1} = \theta_n + \frac{\Sigma}{2} \hat{\nabla}^{y_{\theta_n}} \log \pi(\theta_n|y) + C\eta$  ,  $\eta \sim N(0, \Sigma)$ .
end for

```

Note that a similar algorithm has been proposed in (Welling and Teh 2011; Ahn, Korattikara and Welling 2012) in the context of big data situations, where the gradient of the logarithm of the target distribution is estimated using mini-batches of the data.

We let \hat{P}_Σ denote the corresponding transition kernel arising from Algorithm 4. We now prove that the Stochastic gradient Langevin algorithm, (Algorithm 4), will converge to the discrete-time Langevin diffusion with transition kernel resulting from Algorithm 3.

2.3 Towards theoretical guarantees for the noisy Langevin algorithm

In this case, the approximation guarantees are not as clear as they are for the noisy Metropolis-Hastings algorithm. To begin, there are two levels of approximation:

- the transition kernel P_Σ targets a distribution π_Σ that might be far away from $\pi(\cdot|y)$.
- Moreover, one does not simulate at each step from P_Σ but rather from \hat{P}_Σ .

The first point requires one to control the distance between π_Σ and $\pi(\cdot|y)$. Such an analysis is possible. Here we refer the reader to Proposition 1 in (Dalalyan and Tsybakov 2012) and also to Roberts and Stramer (Roberts and Stramer 2002) for different discretization schemes. For the second point, we believe unfortunately that Theorem 2.1 is not enough, as we now explain. It is possible to control $\|\hat{P}_\Sigma - P_\Sigma\|$ as Lemma 2.3 illustrates.

Lemma 2.3

$$\|P_\Sigma - \hat{P}_\Sigma\| \leq \sqrt{\frac{\delta}{2}}$$

where

$$\delta = \mathbb{E}_{y_{\theta_n} \sim F_{\theta_n}} \left\{ \exp \left[\frac{1}{2} \left\| \Sigma^{\frac{1}{2}} (\nabla \log \pi(\theta_n) - \hat{\nabla}^{y_{\theta_n}} \log \pi(\theta_n)) \right\|^2 \right] - 1 \right\}.$$

However, Roberts and Tweedie (1996a) show that in many situations, the chain generated P_Σ is not even geometrically ergodic. They provide special cases where it is geometrically ergodic, but it is not clear that this chain can be uniformly ergodic.

It may be that Theorem 2.1 is not the appropriate tool in this situation. So the theoretical understanding of this noisy Langevin algorithm is still an open question, although we feel that our work here is a good starting point in this direction. We believe that this is a question of major interest - it is also the approach underlying the papers (Welling and Teh 2011; Ahn, Korattikara and Welling 2012). This will be the focus of our future work.

2.4 Connection with the pseudo-marginal approach

There is a clear connection between this paper and the pseudo-marginal approaches described in (Beaumont 2003) and (Andrieu and Roberts 2009). In both cases a noisy acceptance probability is considered, but in pseudo-marginal approaches this is a consequence of using an estimate of the desired target distribution at each θ , rather than the true value. Before proceeding further, we make precise some of the terminology used in (Beaumont 2003) and (Andrieu and Roberts 2009). These papers describe two alternative algorithms, the “Monte Carlo within Metropolis” (MCWM) approach, and “grouped independence MH” (GIMH). In both cases an unbiased importance sampling estimator, $\hat{\pi}$, is used in place of the desired target π , however the overall algorithms proceed slightly differently. The $(i + 1)$ th iteration of the MCWM algorithm is shown in algorithm 5.

GIMH differs from MCWM as follows. In MCWM the estimate of the target in the denominator is recomputed at every iteration of the MCMC, whereas in GIMH it is reused from the previous iteration. The property that is the focus of (Andrieu and Roberts 2009) is that GIMH actually has the desired target distribution π - this can be seen by viewing the algorithm as an MCMC algorithm targeting an extended target distribution including the auxiliary variables. The same argument holds when using any unbiased estimator of the target. As regards our focus in this paper, GIMH is something of a special case, and our framework has more in common with MCWM. We note that despite its exactness, there is

Algorithm 5 MCWM

for $n = 0$ to I **do**

 Draw $\theta' \sim h(\cdot|\theta_n)$.

 Draw $z' \sim G(\cdot|\theta')$, $z \sim G(\cdot|\theta)$, where G is an importance proposal and z' and z are random vectors of size N .

 Calculate the acceptance probability, $\alpha(\theta_n, \theta')$, where $\hat{\pi}_z^N$ and $\hat{\pi}_{z'}^N$ denote the importance sampling approximation to π based on auxiliary variables z and z' respectively:

 Set $\theta_{n+1} = \theta'$ with probability $\min(1, \hat{\alpha}(\theta', \theta_n))$, where

$$\hat{\alpha}(\theta', \theta_n) = \frac{\hat{\pi}_{z'}^N(\theta')h(\theta_n|\theta')}{\hat{\pi}_z^N(\theta_n)h(\theta'|\theta_n)},$$

 Otherwise, set $\theta_{n+1} = \theta_n$.

end for

no particular reason for estimators from GIMH to be more statistically efficient than those from MCWM.

For our framework to include MCWM as a special case, we require that the distribution $F(\cdot|\theta')$ of the auxiliary variables y' that we use in order to find $\hat{\alpha}(\theta'|\theta, y')$ also needs to depend on θ , so from here on we use $F(\cdot|\theta, \theta')$. For MCWM we have $y' = (z, z')$, with $F(y'|\theta, \theta') = G(z|\theta)G(z'|\theta')$. We note that this additional dependence only requires minor alterations to Corollary 2.2 and its proof. Corollary 2.2 and its proof share some characteristics with the special case (Andrieu and Roberts 2009) where they show that there always exists an N such that an arbitrarily small accuracy can be achieved in the bound for the total variation between the invariant distribution of MCWM (if it exists) and the true target. The arguments in this paper are more general in the sense that the noisy acceptance probability framework covers a larger set of situations but also in that, as we see below, it is sometimes possible to obtain a rate of approximation in terms of N , which in our case is the number of auxiliary variables used in the approximation.

3 Examples

3.1 Gibbs Random Fields

Gibbs random fields (or discrete Markov random fields) are widely used to model complex dependency structure jointly in graphical models in areas including spatial statistics and network analysis. Let $y = \{y_1, \dots, y_M\}$ denote realised data defined on a set of nodes $\{1, \dots, M\}$ of a graph, where each observed value y_i takes values from some finite state

space. The likelihood of y given a vector of parameters $\theta = (\theta_1, \dots, \theta_m)$ is defined as

$$f(y|\theta) \propto \exp(\theta^T s(y)) := q_\theta(y), \quad (3)$$

where $s(y) = (s_1(y), \dots, s_m(y))$ is a vector of statistics which are sufficient for the likelihood. We will use the notation $\mathcal{S} = \sup_{y \in Y} \|s(y)\|$. The constant of proportionality in (3),

$$Z(\theta) = \sum_{y \in Y} \exp(\theta^T s(y)),$$

depends on the parameters θ , and is a summation over all possible realisation of the Gibbs random field. Clearly, $Z(\theta)$ is intractable for all but trivially small situations. The parameter of interest for the Gibbs distribution is θ . Due to the intractability of the normalising constant $Z(\theta)$, inference on θ is problematic. Here and for the remainder of this article we focus on the posterior distribution

$$\pi(\theta|y) \propto \frac{q_\theta(y)}{Z(\theta)} \pi(\theta),$$

where $\pi(\theta)$ denotes the prior distribution for θ . For example, a naive application of the Metropolis-Hastings algorithm when proposing to move from θ_i to $\theta' \sim h(\cdot|\theta_i)$ results in the acceptance probability,

$$\alpha(\theta', \theta) = \min \left(1, \frac{q_{\theta'}(y) \pi(\theta') h(\theta|\theta')}{q_\theta(y) \pi(\theta) h(\theta'|\theta)} \times \frac{Z(\theta)}{Z(\theta')} \right), \quad (4)$$

depending on the intractable ratio $\frac{Z(\theta)}{Z(\theta')}$.

One method to overcome this computational bottleneck is to use an approximation of the likelihood $f(y|\theta)$. A composite likelihood approximation of the true likelihood, such as that of (Besag 1974), is most commonly used. This approximation consists of a product of easily normalised full-conditional distributions. The most basic composite likelihood is the pseudo likelihood which comprised of the product of full-conditional distributions of each y_i ,

$$f(y|\theta) \approx \prod_{i=1}^M f(y_i|y_{-i}, \theta).$$

However this approximation of the true likelihood can give unreliable estimates of θ (Friel and Pettitt 2004), (Friel *et al* 2009).

3.2 Exchange Algorithm

A more sophisticated approach is to use the Exchange algorithm. Murray *et al.* (2006) extended the work of Møller *et al.* (2006) to allow inference on doubly intractable distributions using the exchange algorithm. The algorithm samples from an augmented distribution

$$\pi(\theta', y', \theta|y) \propto f(y|\theta) \pi(\theta) h(\theta'|\theta) f(y'|\theta')$$

whose marginal distribution for θ is the posterior of interest. Here the auxiliary distribution $f(y'|\theta')$ is the same likelihood model in which y is defined. By sampling from this augmented distribution, the acceptance formula simplifies, as can be seen in algorithm 6, where the normalising constants arising from the likelihood and auxiliary likelihood cancel. One

Algorithm 6 Exchange algorithm

for $n = 0$ to I **do**

 Draw $\theta' \sim h(\cdot|\theta_n)$.

 Draw $y' \sim f(\cdot|\theta')$.

 Set $\theta_{n+1} = \theta'$ with probability $\min(1, \alpha(\theta', \theta_n, y'))$, where

$$\alpha(\theta', \theta_n, y') = \frac{q_{\theta'}(y)\pi(\theta')h(\theta_n|\theta')q_{\theta_n}(y')}{q_{\theta_n}(y)\pi(\theta_n)h(\theta'|\theta_n)q_{\theta'}(y')} \times \frac{Z(\theta_n)Z(\theta')}{Z(\theta')Z(\theta_n)},$$

 Otherwise, set $\theta_{n+1} = \theta_n$.

end for

difficulty of implementing the exchange algorithm is the requirement to sample $y' \sim f(\cdot|\theta')$, perfect sampling (Propp and Wilson 1996) is often possible for Markov random field models. However when the exchange algorithm is used with MRFs the resultant chains may not mix well. For example, Caimo and Friel (2011) used adaptive direction sampling (Gilks, Roberts and George 1994) to improve the mixing of the exchange algorithm when used with ERGM models.

Murray *et al.* (2006) proposed the following interpretation of the exchange algorithm. If we compare the acceptance ratios in the M-H and Exchange algorithm, the only difference is that the ratio of the normalising constants in the M-H acceptance probability $Z(\theta)/Z(\theta')$ is replaced by $q_\theta(y')/q_{\theta'}(y')$ in the exchange probability. This ratio of un-normalised likelihoods is in fact an unbiased importance sampling estimator of the ratio of normalising constants since it holds that

$$\mathbb{E}_{y' \sim f(\cdot|\theta')} \left(\frac{q_\theta(y')}{q_{\theta'}(y')} \right) = \frac{Z(\theta)}{Z(\theta')}. \quad (5)$$

A natural extension is therefore to use a better unbiased estimator of $Z(\theta)/Z(\theta')$ at each step of the exchange algorithm. At each step we could simulate a number of auxiliary variables (y'_1, \dots, y'_N) from $f(\cdot|\theta)$, then approximate the ratio of normalising constants by

$$\frac{1}{N} \sum_{i=1}^N \frac{q_\theta(y'_i)}{q_{\theta'}(y'_i)} \approx \frac{Z(\theta)}{Z(\theta')}. \quad (6)$$

3.3 Noisy exchange algorithm

Algorithm 7 results from using an importance sampling estimator of intractable ratio of normalising constants following (6). We term this algorithm the noisy exchange algorithm.

In particular, note that the acceptance ratio is replaced by an estimate $\hat{\alpha}$. Note further that when $N = 1$ this will be equivalent to the exchange algorithm, and when $N \rightarrow \infty$ this will be equivalent to the standard Metropolis Hastings algorithm. Both of these algorithms leave the target posterior invariant. However when $1 < N < \infty$ this algorithm is not guaranteed to sample from the posterior.

Algorithm 7 Noisy Exchange algorithm

for $n = 0$ to I **do**

 Draw $\theta' \sim h(\cdot|\theta_n)$.

for $i = 1$ to N **do**

 Draw $y'_i \sim f(\cdot|\theta')$.

end for

 Define $y_{\theta'} = \{y'_1, \dots, y'_N\}$

 Set $\theta_{n+1} = \theta'$ with probability $\min(1, \hat{\alpha}(\theta', \theta_n, y_{\theta'}))$, where

$$\hat{\alpha}(\theta', \theta_n, y_{\theta'}) = \frac{q_{\theta'}(y)\pi(\theta')h(\theta_n|\theta')}{q_{\theta_n}(y)\pi(\theta_n)h(\theta'|\theta_n)} \frac{1}{N} \sum_{i=1}^N \frac{q_{\theta_n}(y'_i)}{q_{\theta'}(y'_i)}.$$

 Otherwise, set $\theta_{n+1} = \theta_n$.

end for

We will now show that under certain assumptions, as $N \rightarrow \infty$ the noisy exchange exchange algorithm will yield a Markov chain which will converge to the target posterior density. To do so, we can apply Lemma 2.2. First, we define some notation and assumptions that will be used to prove this Lemma.

(A1) there is a constant c_π such that $1/c_\pi \leq \pi(\theta) \leq c_\pi$.

(A2) there is a constant c_h such that $1/c_h \leq h(\theta'|\theta) \leq c_h$.

(A3) for any θ and θ' in Θ ,

$$\text{Var}_{y' \sim f(y'|\theta')} \left(\frac{q_{\theta_n}(y')}{q_{\theta'}(y')} \right) < +\infty.$$

Note that when (A1) or (A2) is satisfied, we necessarily have that Θ is a bounded set, in this case, we put $T = \sup_{\theta \in \Theta} \|\theta\|$. This also means that $0 < \exp(-T\mathcal{S}) \leq q_\theta(y) \leq \exp(T\mathcal{S})$ for any θ and \mathcal{S} , we then put $\mathcal{K} := \exp(T\mathcal{S})$. Also, note that this immediately implies Assumption (A3) because in this case, $\text{Var}_{y' \sim f(y'|\theta')} (q_{\theta_n}(y')/q_{\theta'}(y')) \leq \mathcal{K}^2$, so Assumption (A3) is weaker than (A1) and than (A2).

Lemma 3.1 Under (A3), $\hat{a}(\theta'|\theta, y')$ satisfies (H2) in Lemma 2.2 with

$$\begin{aligned} \mathbb{E}_{y' \sim f(\cdot|\theta')} |\hat{a}(\theta, \theta', y') - a(\theta, \theta')| &\leq \delta(\theta, \theta') \\ &= \frac{1}{\sqrt{N}} \frac{h(\theta|\theta')\pi(\theta')q_{\theta'}(y)}{h(\theta'|\theta)\pi(\theta)q_{\theta}(y)} \sqrt{\text{Var}_{y' \sim f(y'|\theta')} \left(\frac{q_{\theta_n}(y')}{q_{\theta'}(y')} \right)}. \end{aligned}$$

Theorem 3.2 Under (A1) and (A2) then (H2) in Lemma 2.2 is satisfied with

$$\delta(\theta, \theta') \leq \frac{c_h^2 c_\pi^2 \mathcal{K}^4}{\sqrt{N}},$$

and

$$\sup_{\theta_0 \in \Theta} \|\delta_{\theta_0} P^n - \delta_{\theta_0} \hat{P}^n\| \leq \frac{\mathcal{C}}{\sqrt{N}}$$

where $\mathcal{C} = \mathcal{C}(c_\pi, c_h, \mathcal{K})$ is explicitly known.

Note that Liang and Jin (2011) presents a similar algorithm to that above. However in contrast to Lemma 3.1, the results in (Liang and Jin 2011) do not explicitly provide a rate of approximation with respect to N . Lemma 2.2, page 9 in (Liang and Jin 2011) only states that there exists a N large enough to reach arbitrarily small accuracy $\epsilon > 0$.

3.4 Noisy Langevin algorithm for Gibbs random fields

The discrete-time Langevin approximation (3) is unavailable for Gibbs random fields since the gradient of the log posterior, $\nabla \log \pi(\theta_i|y)$ is analytically intractable, in general. However Algorithm 4 can be used using a Monte Carlo estimate of the gradient, as follows.

$$\begin{aligned} \log(\pi(\theta|y)) &= \theta^T s(y) - \log(z(\theta)) + \log \pi(\theta) - \log(\pi(y)) \\ \nabla \log(\pi(\theta|y)) &= s(y) - \frac{z'(\theta)}{z(\theta)} + \nabla \log \pi(\theta) \\ &= s(y) - \frac{\sum s(y) [\exp \theta^T s(y)]}{\sum \exp(\theta^T s(y))} + \nabla \log \pi(\theta) \\ &= s(y) - \mathbb{E}_{y|\theta}[s(y)] + \nabla \log \pi(\theta) \end{aligned} \tag{7}$$

In practice, $\mathbb{E}_{y' \sim f_\theta}[s(y')]$ is usually not known - an exact evaluation of this quantity would require an evaluation of $Z(\theta)$. However, it is possible to estimate it through Monte-Carlo simulations. If we simulate $y_\theta = (y'_1, \dots, y'_n) \sim f(\cdot|\theta)$, then $\mathbb{E}_{y|\theta}[s(y)]$ can be estimated using $\sum_i^n s(y'_i)/n$. This gives an estimate of the gradient at θ from (7).

$$\widehat{\nabla}^{y_\theta} \log \pi(\theta|y) = s(y) - \frac{1}{N} \sum_i^N s(y'_i) + \nabla \log \pi(\theta).$$

In turn this yield the following noisy discretized Langevin algorithm.

Remark that in this case, the bound in Lemma 2.3 can be evaluated.

Algorithm 8 Noisy discretized Langevin algorithm for Gibbs random fields

```

for  $n = 0$  to  $I$  do
  for  $i = 1$  to  $N$  do
    Draw  $y'_i \sim f(\cdot | \theta_n)$ .
  end for
  Define  $y_{\theta_n} = \{y'_1, \dots, y'_N\}$ ,

  Calculate  $\widehat{\nabla}^{y_{\theta_n}} \log \pi(\theta_n | y) = \nabla \log \pi(\theta_n) + s(y) - \frac{1}{N} \sum_{i=1}^N s(y'_i)$ .

  Set
    
$$\theta_{n+1} = \theta_n + \frac{\Sigma}{2} \widehat{\nabla}^{y_{\theta_n}} \log \pi(\theta_n | y) + \eta_n, \quad \text{where } \eta_n \text{ are i.i.d. } \mathcal{N}(0, \Sigma).$$

  end for

```

Lemma 3.3 *As soon as $N > 4k\mathcal{S}^2\|\Sigma\|^2$, the δ in Lemma 2.3 is finite with*

$$\delta = \exp\left(\frac{k \log(N)}{4\mathcal{S}^2\|\Sigma\|^2 N}\right) - 1 + \frac{4k\sqrt{\pi}\mathcal{S}\|\Sigma\|}{N} \sim_{N \rightarrow \infty} \frac{k \log\left(\frac{N}{k}\right)}{4\mathcal{S}^2\|\Sigma\|^2 N}$$

(where $\|\Sigma\| = \sup\{\|\Sigma x\|, \|x\| = 1\}$).

Remark 3.1 *We remark again that this lemma is unfortunately not enough to provide guarantees for the noisy Langevin algorithm: usually, the kernel P_Σ does not generate a uniformly ergodic chain, which is a key requirement in our framework. However, although insufficient, using Theorem 1 page 352 in (Roberts and Tweedie 1996a), we can provide situations where the chain is geometrically ergodic. E.g., consider the case $\theta \in \mathbb{R}$ with a continuously differentiable prior $\pi(\theta)$ with $\nabla \log \pi(\theta) \asymp -c\theta^d$ when $\theta \rightarrow \infty$ for some $d \in (0, 1]$ and $\nabla \log \pi(\theta) \asymp c|\theta|^d$ when $\theta \rightarrow -\infty$. In this situation, the chain is geometrically ergodic in all cases when $d < 1$, and for $\Sigma^2 < \frac{2}{c}$ in the case where $d = 1$.*

3.5 MALA-exchange

An approach to ensure that the Markov chain from Algorithm 8 targets the true density, is to include an accept/reject step at each iteration in this algorithm using a Metropolis adjusted Langevin (MALA) correction. We adapt the Exchange algorithm using this proposal, yielding Algorithm 9.

The accept/reject step ensures that the distribution targets the correct posterior density. If the stochastic gradient $\widehat{\nabla}$ approximates the true gradient well, then the proposal value at each iteration should be guided towards areas of high density. This will allow the algorithm to explore the posterior more efficiently when compared with a random walk proposal.

Algorithm 9 MALA-exchange

Initialise; set Σ ,

for $i = 1$ to N **do**

 Draw $y_i \sim f(\cdot|\theta_0)$.

end for

Define $y_{\theta_0} = \{y_1, \dots, y_N\}$,

Calculate $\widehat{\nabla}^{y_{\theta_0}} \log \pi(\theta_0|y) = \nabla \log \pi(\theta_0) + s(y) - \frac{1}{N} \sum_{i=1}^N s(y_i)$.

for $n = 0$ to I **do**

 Draw $\theta' = \theta_n + \frac{\Sigma}{2} \widehat{\nabla}^{y_{\theta_n}} \log \pi(\theta_n|y) + \eta$, $\eta \sim N(0, \Sigma)$.

for $i = 1$ to N **do**

 Draw $y'_i \sim f(\cdot|\theta')$.

end for

 Define $y_{\theta'} = \{y'_1, \dots, y'_N\}$.

 Calculate $\widehat{\nabla}^{y_{\theta'}} \log \pi(\theta'|y) = \nabla \log \pi(\theta') + s(y) - \frac{1}{N} \sum_{i=1}^N s(y'_i)$.

 Set $\theta_{n+1} = \theta'$ and $y_{\theta_{n+1}} = y_{\theta'}$ with probability $\min(1, \alpha(\theta', \theta_n, y_{\theta_n}))$,

 where $\alpha(\theta', \theta_n, y_{\theta_n}) = \frac{q_{\theta'}(y)\pi(\theta')h(\theta_n|\theta', y_{\theta'})q_{\theta_n}(y'_1)}{q_{\theta_n}(y)\pi(\theta_n)h(\theta'|\theta_n, y_{\theta_n})q_{\theta'}(y'_1)}$,

 and $h(\theta_n|\theta', y_{\theta'}) \sim N\left(\theta' + \widehat{\nabla}^{y_{\theta'}} \log \pi(\theta'|y), \Sigma\right)$.

 Otherwise, set $\theta_{n+1} = \theta_n$ and $y_{\theta_{n+1}} = y_{\theta_n}$.

end for

3.6 Noisy MALA-exchange

In an approach identical to that in Section 3.3 one could view the ratio $q_{\theta_i}(y')/q_{\theta'}(y')$ in the acceptance ratio from Algorithm 9 as an importance sampling estimator of $Z(\theta')/Z(\theta_i)$. This suggests that one could replace this ratio of un-normalised densities with a Monte Carlo estimator using draws from $f(y|\theta')$, as described in (6). Here, we suggest that the draws used to estimate the log gradient could serve this purpose. This yields the noisy MALA-exchange algorithm which we outline below.

Algorithm 10 noisy MALA-exchange

```

Initialise; set  $\Sigma$ ,
for  $i = 1$  to  $N$  do
    Draw  $y_i \sim f(\cdot|\theta_0)$ .
end for
Define  $y_{\theta_0} = \{y_1, \dots, y_N\}$ ,
Calculate  $\widehat{\nabla}^{y_{\theta_0}} \log \pi(\theta_0|y) = \nabla \log \pi(\theta_0) + s(y) - \frac{1}{N} \sum_{i=1}^N s(y_i)$ .

for  $n = 0$  to  $I$  do
    Draw  $\theta' = \theta_n + \frac{\Sigma}{2} \widehat{\nabla}^{y_{\theta_n}} \log \pi(\theta_n|y) + \eta \quad \eta \sim N(0, \Sigma)$ .

    for  $i = 1$  to  $N$  do
        Draw  $y'_i \sim f(\cdot|\theta')$ .
    end for
    define  $y_{\theta'} = \{y'_1, \dots, y'_N\}$ .

    Calculate  $\widehat{\nabla}^{y_{\theta'}} \log \pi(\theta'|y) = \nabla \log \pi(\theta') + s(y) - \frac{1}{N} \sum_{i=1}^N s(y'_i)$ .

    Set  $\theta_{n+1} = \theta'$  and  $y_{\theta_{n+1}} = y_{\theta'}$  with probability  $\min(1, \hat{\alpha}(\theta', \theta_n, y_{\theta_n}))$ 
    where  $\hat{\alpha}(\theta', \theta_n, y_{\theta_n}) = \frac{q_{\theta'}(y)\pi(\theta')h(\theta_n|\theta', y'_{\theta_n})}{q_{\theta_n}(y)\pi(\theta_n)h(\theta'|\theta_n, y'_{\theta_n})} \frac{1}{N} \sum_{i=1}^N \frac{q_{\theta_n}(y'_i)}{q_{\theta'}(y'_i)}$ ,
    and  $h(\theta_n|\theta', y_{\theta'}) \sim N\left(\theta' + \widehat{\nabla}^{y_{\theta'}} \log \pi(\theta'|y), \Sigma\right)$ .

    Otherwise, set  $\theta_{n+1} = \theta_n$  and  $y_{\theta_{n+1}} = y_{\theta_n}$ .
end for

```

4 Experiments

We first demonstrate our algorithms on a simple single parameter model, the Ising model and then apply our methodology to some challenging models for the analysis of network data.

4.1 Ising study

The Ising model is defined on a rectangular lattice or grid. It is used to model the spatial distribution of binary variables, taking values -1 and 1 . The joint density of the Ising model can be written as

$$f(y|\theta) = \frac{1}{Z(\theta)} \exp \left\{ \theta \sum_{j=1}^M \sum_{i \sim j} y_i y_j \right\}$$

where $i \sim j$ denotes that i and j are neighbours and $Z(\theta) = \sum_{\mathbf{y}} \exp \left\{ \theta \sum_{j=1}^M \sum_{i \sim j} y_i y_j \right\}$.

The normalising constant $Z(\theta)$ is rarely available analytically since this relies on taking the summation over all different possible realisations of the lattice. For a lattice with M nodes this equates to $2^{\frac{M(M-1)}{2}}$ different possible lattice formations.

For our study, we simulated 20 grids of size 16×16 . This size lattice is sufficiently small enough such that the normalising constant $Z(\theta)$ can be calculated exactly (36.5 minutes for each graph) using a recursive forward-backward algorithm (Reeves and Pettitt 2004; Friel and Rue 2007), giving a gold standard with which to compare the other algorithms. This is done by calculating the exact density over a fine grid of θ values, $\{\theta_1, \theta_I\}$ over the interval $[-0.4, 0.8]$, which cover the effective range of values that θ can take. We normalise $\pi(\theta_i|y)$ by numerically integrating over the un-normalised density.

$$\hat{\pi}(y) = \sum_{i=2}^I \frac{(\theta_i - \theta_{i-1})}{2} \left[\frac{q_{\theta_i}(y)}{Z(\theta_i)} \pi(\theta_i) + \frac{q_{\theta_{i-1}}(y)}{Z(\theta_{i-1})} \pi(\theta_{i-1}) \right], \quad (8)$$

yielding

$$\pi(\theta_i|y) \approx \frac{q_{\theta_i}(y)}{Z(\theta_i)} \frac{\pi(\theta_i)}{\hat{\pi}(y)}.$$

Each of the algorithms was run for 30 seconds on each of the 20 datasets, at each iteration the auxiliary step to draw y' was run for 1000 iterations. For each of the noisy, Langevin and MALA exchange, an extra $N = 100$ draws were taken during the auxiliary step to use as the simulated graphs $y_{\theta'}$.

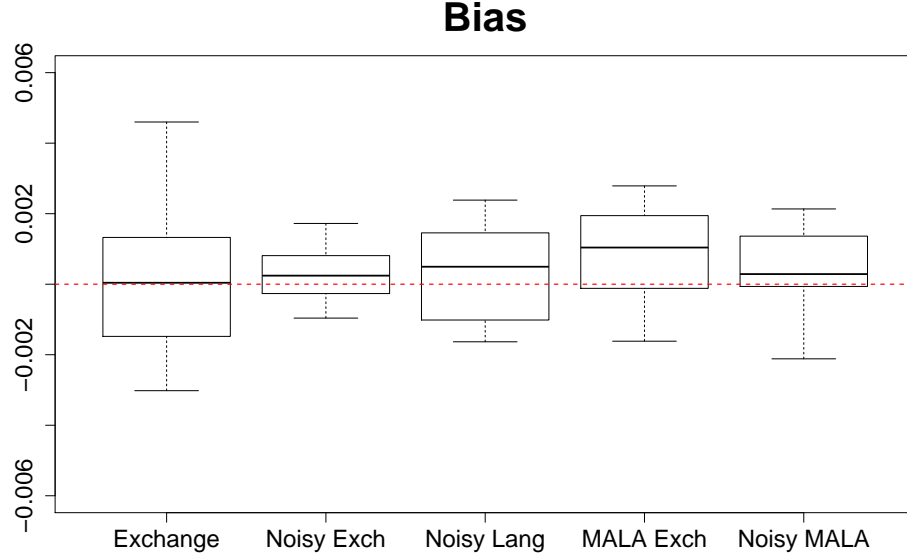


Figure 1: Boxplot of the bias estimate of θ for 20 datasets corresponding to the exchange, importance sampling exchange, Langevin and MALA algorithms.

Figure 1 shows the bias of the posterior means for each of the algorithms. We see that both the noisy exchange algorithm and the Langevin algorithm have a much smaller bias when compared to the two exchange algorithms. The two noisy algorithms perform better than the two exact algorithms. This is due to the improved mixing in the approximate algorithms, even though the true distribution is only approximately targeted. There is a trade off here between the bias and the efficiency. As the step size decreases, both the efficiency and bias decrease. The MALA-exchange appears better than the exchange, this is due to the informed proposal used in the MALA algorithm $\hat{\nabla} \log \pi(\theta|y)$. This informed proposal means the MALA-exchange will target areas of high probability in the posterior density, therefore increasing the chances of accepting a move at each iteration when compared to the standard exchange. Finally, in Figure 2 we display the estimated posterior density for each of the five algorithms together with the true posterior density for one of the 20 datasets in the simulation study.

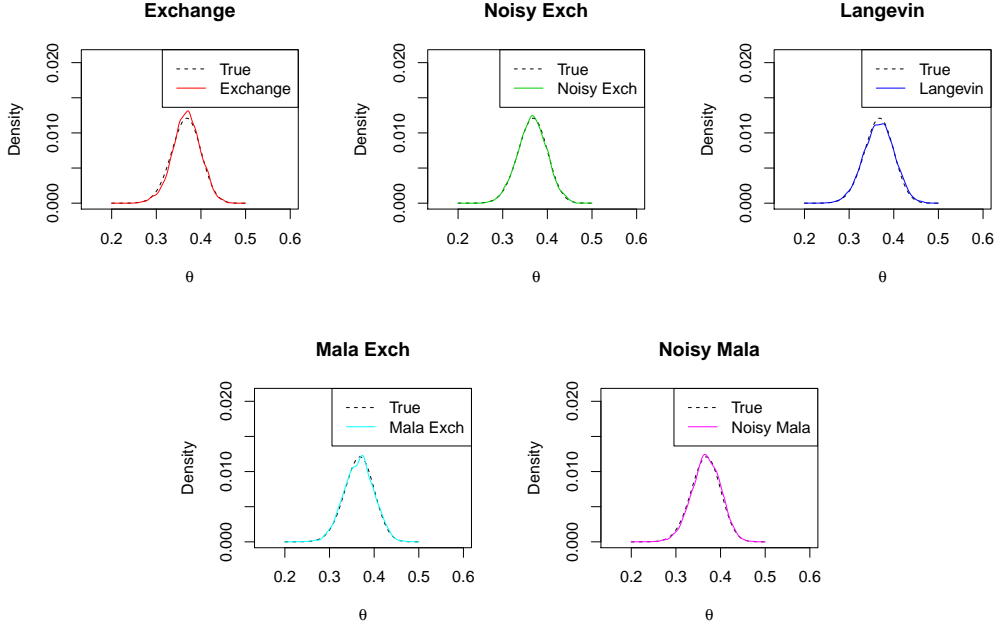


Figure 2: Estimated posterior densities corresponding to the exact and noisy algorithms corresponding to one of the datasets used in the Ising simulation study.

4.2 ERGM study

Here we explore how our algorithms may be applied to the exponential random graph model (ERGM) (Robins *et al* 2007) which is widely used in social network analysis. An ERGM is defined on a random adjacency matrix \mathbf{Y} of a graph on n nodes (or actors) and a set of edges (dyadic relationships) $\{Y_{ij} : i = 1, \dots, M; j = 1, \dots, M\}$ where $Y_{ij} = 1$ if the pair (i, j) is connected by an edge, and $Y_{ij} = 0$ otherwise. An edge connecting a node to itself is not permitted so $Y_{ii} = 0$. The dyadic variables maybe be undirected, whereby $Y_{ij} = Y_{ji}$ for each pair (i, j) , or directed, whereby a directed edge from node i to node j is not necessarily reciprocated.

The likelihood of an observed network y is modelled in terms of a collection of sufficient statistics $\{s_1(y), \dots, s_m(y)\}$, each with corresponding parameter vector $\theta = \{\theta_1, \dots, \theta_m\}$,

$$f(y|\theta) = \frac{q_\theta(y)}{Z(\theta)} = \frac{\exp\{\sum_{l=1}^m \theta_l s_l(y)\}}{Z(\theta)}.$$

For example, typical statistics include $s_1(y) = \sum_{i < j} y_{ij}$ and $s_2(y) = \sum_{i < j < k} y_{ik} y_{jk}$ which are, respectively, the observed number of edges and two-stars, that is, the number of configurations of pairs of edges which share a common node. It is also possible to consider statistics which count the number of triangle configurations, that is, the number of configurations in which nodes i, j, k are all connected to each other.

4.2.1 The Florentine Business dataset

Here, we consider a simple 16 node undirected graph: the Florentine family business graph. This concerns the business relations between some Florentine families in around 1430. The network is displayed in Figure 3. We propose to estimate the following 2-dimensional model.

$$f(y|\theta) = \frac{1}{Z(\theta)} \exp(\theta_1 s_1(y) + \theta_2 s_2(y)),$$

where $s_1(y)$ is the number of edges in the graph and $s_2(y)$ is the number of two-stars.

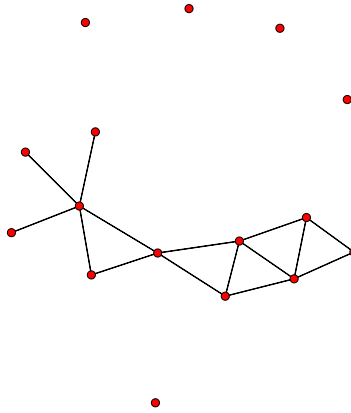


Figure 3: Florentine family business.

Before we could run the algorithms, certain parameters had to be tuned. We used a flat prior $N(0, 100)$ in all of the algorithms. The Langevin, MALA exchange and noisy MALA exchange algorithms all depend on a stepsize matrix Σ . This matrix determines the scale of proposal values for each of the parameters. This matrix should be set up so that proposed values for θ accommodate the different scales of the posterior density of θ . In order to have good mixing in the algorithms we chose a Σ which relates to the shape of the posterior density. Our approach was to aim to relate Σ to the covariance of the posterior density. To do this, we equated Σ to an estimate of the inverse of the second derivative of the log posterior at the *maximum a posteriori* estimate θ^* . As the true value of the MAP is unknown, we used a Robbins-Monro algorithm (Robbins and Monro 1951) to estimate this. The Robbins-Monro algorithm takes steps in the direction of the slope of the distribution. It is very similar to Algorithm 8 except without the added noise and follows the stochastic process

$$\theta_{n+1} = \theta_n + \epsilon_n \widehat{\nabla}^{y_{\theta_n}} \log \pi(\theta_n | y),$$

where $\sum_{i=0}^N \epsilon_n < \infty$ and $\sum_{i=0}^N \epsilon_n^2 < \infty$.

The values of ϵ decrease over time and once the difference between successive values of this process is less than a specified tolerance level, the algorithm is deemed to have converged to the MAP. The second derivative of the log posterior is derived by differentiating (7) yielding

$$\nabla^2 \log \pi(\theta^*|y) = \text{Cov}_{y^*|\theta^*}(s(y^*)) + \nabla^2 \log \pi(\theta^*) \quad (9)$$

In turn, $\text{Cov}_{y^*|\theta^*}(s(y^*))$ from (9) can be estimated using Monte Carlo based on draws from the likelihood $f(y|\theta^*)$. We used the inverse of the estimate of the second derivative of the log posterior as an estimate for the curvature of our log posterior distribution. The matrix Σ we used was this estimate of the curvature multiplied by a scalar. We multiply by a scalar to achieve different acceptance rates for the algorithms. This is similar to choosing a variance for the proposal in a standard Metropolis-Hastings algorithm. If too small a value is chosen for the scalar, the algorithm will propose small steps and take a long time to fully explore the posterior distribution. If too large a value is chosen for the scalar, the chain will inefficiently explore the target distribution. A number of pilot runs were made to find a value for the scalar which gave the desired acceptance rates for each of the algorithms. The MALA exchange and Noisy MALA exchange algorithms were tuned to have an acceptance rate of approximately 25% and a similar Σ matrix was used in the noisy Langevin algorithm. If the second derivative matrix is singular, a problem can arise, in that is impossible to calculate the inverse of the matrix. Further information on singular matrices can be found in numerical linear algebra literature, such as (Golub and Loan 1996).

The algorithms were time normalised, each using 30 seconds of CPU time. An extra $N = 50$ graphs were simulated for the noisy exchange, noisy Langevin, MALA exchange and noisy MALA exchange algorithms. The auxiliary step to draw y' was run for 1000 iterations followed by an extra 200 iterations thinned by a factor of 4 yielding $N = 50$ graphs. To compare the results to a “ground truth”, the BERGM algorithm of (Caimo and Friel 2011) was run for an large number of iterations equating to 2 hours of CPU time. This algorithm involves a population MCMC algorithm and uses the current state of the population to help make informed proposals for the chains within the population.

Table 1 shows the posterior means and standard deviations for the various algorithms. Figures 4 and 5 shows the chains, densities and autocorrelation plots. In Table 1 we see that the noisy exchange algorithm had improved mean estimates when compared to the exchange algorithm. The MALA exchange and Noisy MALA exchange algorithms both had better mean estimates than the noisy Langevin algorithm, although in all cases the posterior standard deviation was underestimated.

The ACF plots in Figures 4 and 5 show how all of the noisy algorithms displayed better mixing when compared to the exchange algorithm. The density plots show that all of the algorithms with the exception of the noisy Langevin estimated the mode of the true density well but they underestimated the standard deviation.

The noisy Langevin performed poorly. A problem of Langevin diffusion as pointed out in (Girolami and Calderhead 2011) is that convergence to the invariant distribution is no longer guaranteed for finite step size owing to the first-order integration error that is introduced. This discrepancy is corrected by the Metropolis step in the MALA exchange and noisy

MALA exchange but not in the Langevin algorithm. Since our Noisy Langevin algorithm approximates Langevin diffusion we are approximating an approximation. There are two levels of approximations which leaves more room for error.

Method	Edge		2-star	
	Mean	SD	Mean	SD
BERGM	-2.675	0.647	0.188	0.155
Exchange	-2.573	0.568	0.146	0.133
Noisy Exchange	-2.686	0.526	0.167	0.122
Noisy Langevin	-2.281	0.513	0.081	0.119
MALA Exchange	-2.518	0.62	0.136	0.128
Noisy MALA	-2.584	0.498	0.144	0.113

Table 1: Posterior means and standard deviations.

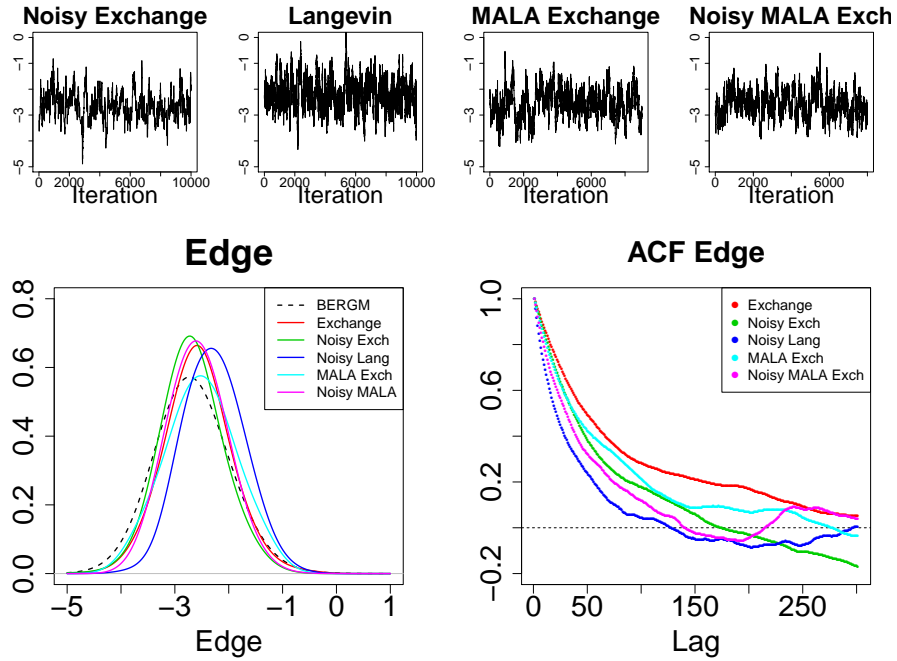


Figure 4: Chains, density plot and ACF plot for the edge statistic.

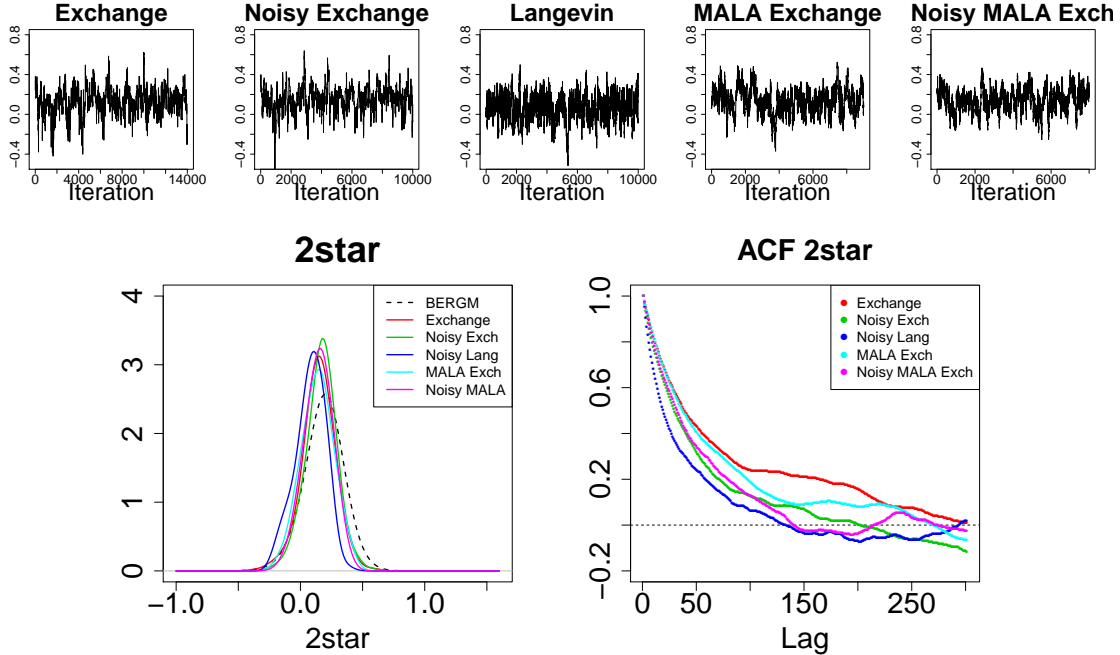


Figure 5: Chains, density plot and ACF plot for the 2-star statistic.

4.2.2 The Molecule dataset

The Molecule dataset is a 20 node graph, shown in Figure 6. We consider a four parameter model which includes the number of edges in the graph, the number of two-stars, the number of three-stars and the number of triangles.

$$f(y|\theta) = \frac{1}{Z(\theta)} \exp(\theta_1 s_1(y) + \theta_2 s_2(y) + \theta_3 s_3(y) + \theta_4 s_4(y))$$

The Σ parameter was chosen in a similar fashion to the Florentine business example. The Robbins-Monro algorithm was run for 20,000 iterations to find an estimate of the MAP, 4,000 graphs were then simulated at the estimated MAP and these were used to calculate an estimate of the second derivative using Equation (9). The matrix Σ was the inverse of this estimate was calculated multiplied by a scalar. The scalar was chosen as a value which achieved the desired acceptance rate, a number of pilot runs were used to get a reasonable value for the scalar. This was carried out for both the MALA exchange and noisy MALA exchange and a similar Σ matrix was used for the noisy Langevin algorithm. The ERGM model for the molecule data is more challenging than the model for the Florentine data due to the extra two parameters.

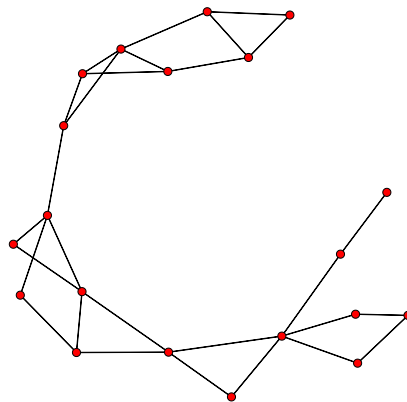


Figure 6: Molecule network

The BERGM algorithm of (Caimo and Friel 2011) was again used as a “ground truth”. This algorithm was run for a large number of iterations equating to 4 hours of CPU time. This gave us accurate estimates against which to compare the various algorithms. The five algorithms were each run for 100 seconds of CPU time. Table 2 shows the posterior mean and standard deviations of each of the four parameters for each of the algorithms. The results for the Molecule dataset model are similar to the Florentine business dataset model. In Table 2 we see that the noisy exchange algorithm improved on the standard exchange algorithm. The MALA exchange improved on noisy Langevin and the Noisy MALA improved on the MALA exchange.

Figure 7 and Figure 8 show the densities and the autocorrelation plots of the algorithms. The autocorrelation plots show that the noisy algorithms had less correlation than the exchange algorithm. The densities show that again the algorithms, when run on the Molecule model, performed in the same manner as the Florentine model. The algorithms with the exception of the noisy Langevin algorithm estimated the mode well but underestimated the standard deviation. The noisy Langevin algorithm did not estimate the mean or standard deviations well.

Method	Edge		2-star		3-Star		Triangle	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
BERGM	2.647	2.754	-1.069	0.953	-0.021	0.483	1.787	0.646
Exchange	1.889	2.142	-0.797	0.744	-0.138	0.385	1.593	0.519
Noisy Exch	1.927	2.444	-0.757	0.823	-0.176	0.422	1.543	0.53
Noisy Lang	1.679	3.65	-0.509	1.429	-0.466	0.787	1.633	0.573
MALA Exch	2.391	2.095	-0.938	0.795	-0.113	0.451	1.454	0.598
Noisy MALA Exch	2.731	2.749	-1.054	0.886	-0.041	0.417	1.519	0.492

Table 2: Posterior means and standard deviations.

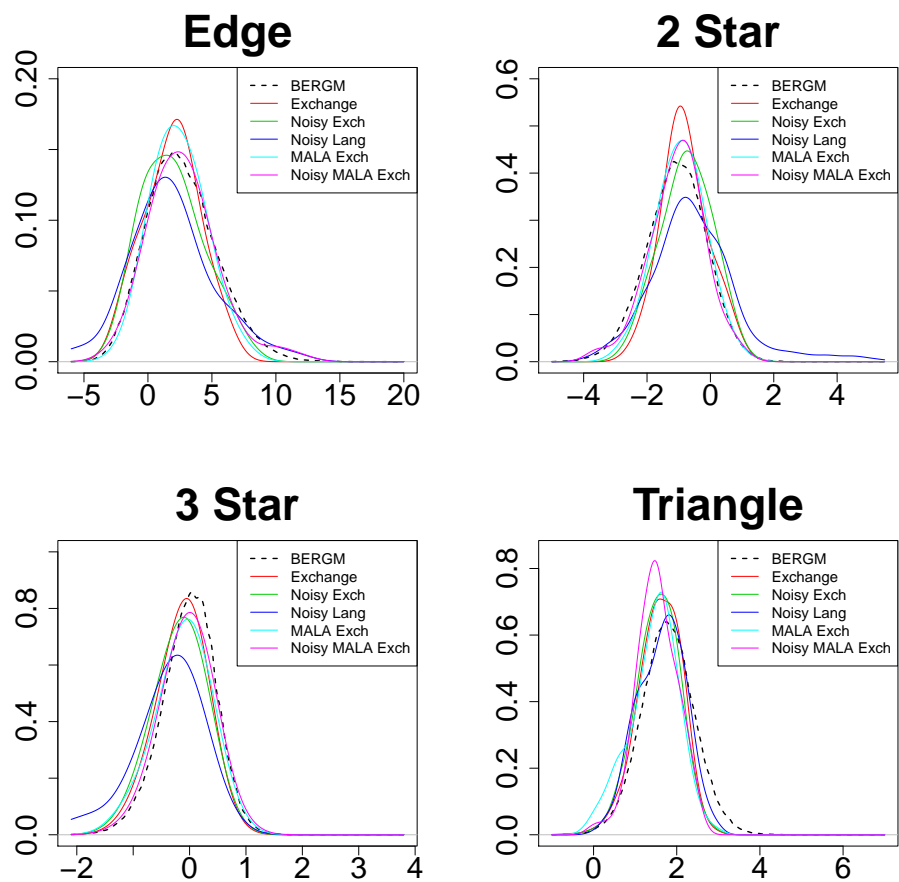


Figure 7: Density plots of the 4 parameters for the molecule example.

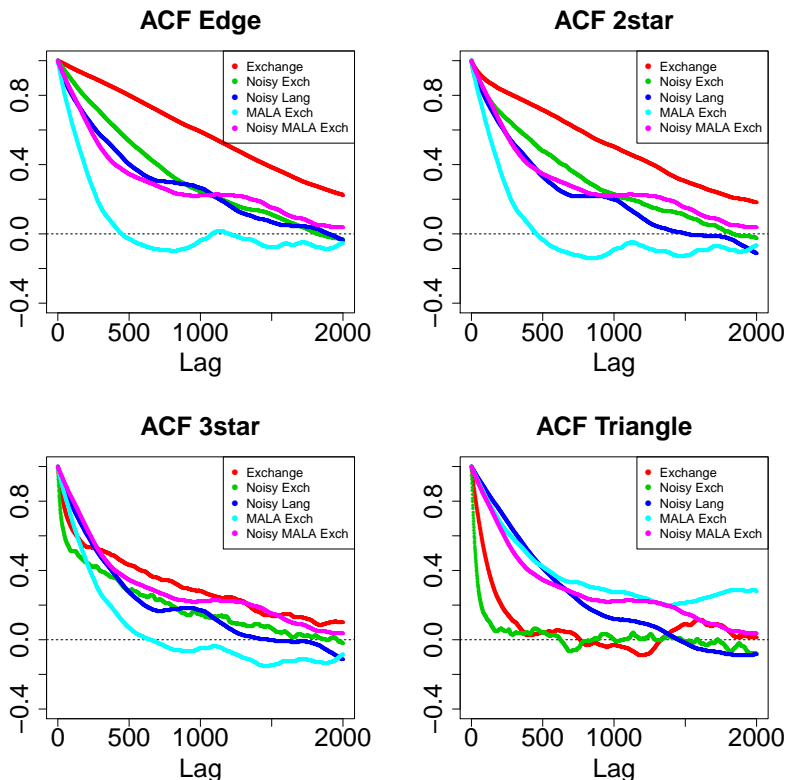


Figure 8: ACF plots for the 4 parameters for the molecule example.

5 Conclusion

The results in this paper give bounds on the total variation between a Markov chain with the desired target distribution, and the Markov chain of a noisy MCMC algorithm. An important question for future work concerns the statistical efficiency of estimators given by ergodic averages of the chain output. This is a key question since the use of noisy MCMC will usually be motivated by the inefficiency of a standard alternative algorithm. This inefficiency may be: statistical, where the standard algorithm is only capable of exploring the parameter space slowly (as can be the case for the standard exchange algorithm); or, computational, where a single iteration of the standard algorithm is too computationally expensive for the method to be practically useful (as is the case for large data sets, examined by Korattikara *et al.* (2014)). If we introduce a noisy MCMC algorithm to overcome the inefficiency, usually the rationale is that the combined statistical and computational efficiency is sufficiently improved to outweigh the effect of any bias that is introduced. To study this theoretically we need to investigate the asymptotic variance of estimators from noisy MCMC algorithms. Andrieu and Vihola (2012) have examined this question for pseudo-marginal algorithms of the GIMH type, and have shown the asymptotic variance for pseudo-marginal algorithms is always larger than for the corresponding “ideal” algorithm. One might expect a similar

result to hold for noisy MCMC algorithms, in which case the effect of this additional variance on top of the aforementioned bias should be a consideration when employing noisy MCMC.

A further area for future work lies in relaxing the requirement for the ideal non-noisy chain to be uniformly ergodic. This property does not hold in many cases: the results in this paper are intended as the first steps towards future work that would obtain results that hold more generally.

Acknowledgements

The Insight Centre for Data Analytics is supported by Science Foundation Ireland under Grant Number SFI/12/RC/2289. Nial Friel’s research was also supported by an Science Foundation Ireland grant: 12/IP/1424.

References

- Ahn, S., A. Korattikara and M. Welling (2012), Bayesian Posterior Sampling via Stochastic Gradient Fisher Scoring. In *Proceedings of the 29th International Conference on Machine Learning*
- Andrieu, C. and G. Roberts (2009), The pseudo-marginal approach for efficient Monte-Carlo computations. *The Annals of Statistics* **37**(2), 697–725
- Andrieu, C. and M. Vihola (2012), Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. Preprint arXiv:1210.1484
- Bardenet, R., A. Doucet and C. Holmes (2014), Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach. In *Proceedings of the 31st International Conference on Machine Learning*
- Beaumont, M. A. (2003), Estimation of population growth or decline in genetically monitored populations. *Genetics* **164**, 1139–1160
- Besag, J. E. (1974), Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society, Series B* **36**, 192–236
- Bottou, L. and O. Bousquet (2011), The tradeoffs of large-scale learning. In S. Sra, S. Nowozin and S. J. Wright (eds.), *Optimization for machine learning*, pp. 351–368, MIT Press
- Bühlmann, P. and S. Van de Geer (2011), *Statistics for high-dimensional data*. Springer
- Caimo, A. and N. Friel (2011), Bayesian inference for exponential random graph models. *Social Networks* **33**, 41–55
- Dalalyan, A. and A. B. Tsybakov (2012), Sparse regression learning by aggregation and Langevin. *J. Comput. System Sci.* **78**(5), 1423–1443
- Friel, N. and A. N. Pettitt (2004), Exact maximum likelihood estimation of the autologistic model on the lattice. *Journal of Computational and Graphical Statistics* **13**, 232–246

- Friel, N., A. N. Pettitt, R. Reeves and E. Wit (2009), Bayesian inference in hidden Markov random fields for binary data defined on large lattices. *Journal of Computational and Graphical Statistics* **18**, 243–261
- Friel, N. and H. Rue (2007), Recursive computing and simulation-free inference for general factorizable models. *Biometrika* **94**, 661–672
- Geman, S. and D. Geman (1984), Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741
- Gilks, W., G. Roberts and E. George (1994), Adaptive direction sampling. *The Statistician* **43**, 179–189
- Girolami, M. and B. Calderhead (2011), Riemann manifold Langevin and Hamiltonian Monte Carlo methods (with discussion). *Journal of the Royal Statistical Society, Series B* **73**, 123–214
- Golub, G. and C. V. Loan (1996), *Matrix computations*. Johns Hopkins University Press, Baltimore, MD
- Kartashov, N. V. (1996), *Strong Stable Markov Chains*. VSP, Utrecht
- Korattikara, A., Y. Chen and M. Welling (2014), Austerity in MCMC Land: Cutting the Metropolis-Hastings Budget. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 681–688
- Liang, F. and I.-H. Jin (2011), An Auxiliary Variables Metropolis-Hastings Algorithm for Sampling from Distributions with Intractable Normalizing Constants. Technical report
- Marin, J.-M., P. Pudlo, C. P. Robert and R. J. Ryder (2012), Approximate Bayesian computational methods. *Statistics and Computing* **22**(6), 1167–1180
- Meyn, S. and R. L. Tweedie (1993), *Markov Chains and Stochastic Stability*. Cambridge University Press
- Mitrophanov, A. Y. (2005), Sensitivity and convergence of uniformly ergodic Markov chains. *Journal of Applied Probability* **42**, 1003–1014
- Moller, J., A. Pettit, R. Reeves and K. Bertheksen (2006), An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika* **93**, 451–458
- Murray, I., Z. Ghahramani and D. MacKay (2006), MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence UAI06*, Arlington, Virginia, AUAI Press
- Nicholls, G. K., C. Fox and A. Muir Watt (2012), Coupled MCMC with a randomized acceptance probability. Preprint arXiv:1205.6857
- Propp, J. and D. Wilson (1996), Exactly sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms* **9**, 223–252

- Reeves, R. and A. N. Pettitt (2004), Efficient recursions for general factorisable models. *Biometrika* **91**, 751–757
- Robbins, H. and S. Monro (1951), A Stochastic Approximation Method. *The Annals of Mathematical Statistics* **22**(3), 400–407
- Roberts, G. O. and O. Stramer (2002), Langevin Diffusions and Metropolis-Hastings Algorithms. *Methodology and Computing in Applied Probability* **4**, 337–357
- Roberts, G. O. and R. L. Tweedie (1996a), Exponential Convergence of Langevin Distributions and their Discrete Approximations. *Bernoulli* **2**(4), 341–363
- Roberts, G. O. and R. L. Tweedie (1996b), Geometric Convergence and Central Limit Theorems for Multidimensional Hastings and Metropolis Algorithm. *Biometrika* **83**(1), 95–110
- Robins, G., P. Pattison, Y. Kalish and D. Lusher (2007), An introduction to exponential random graph models for social networks. *Social Networks* **29**(2), 169–348
- Tibshirani, R. (1996), Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society B* **58**(1), 267–288
- Valiant, L. (1984), A theory of the learnable. *Communications of the ACM* **27**(11), 1134–1142
- Welling, M. and Y. W. Teh (2011), Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 681–688

A Proofs

Proof of Corollary 2.2. We apply Theorem 2.1. First, note that we have

$$P(\theta, d\theta') = \delta_\theta(d\theta') \left[1 - \int dt h(t|\theta) \min(1, \alpha(\theta, t)) \right] + h(\theta'|\theta) \min(1, \alpha(\theta, \theta'))$$

and

$$\begin{aligned} \hat{P}(\theta, d\theta') = \delta_\theta(d\theta') \left[1 - \iint dt dy' h(t|\theta) F_t(y') \min(1, \hat{\alpha}(\theta, t, y')) \right] \\ + \int dy' F_{\theta'}(y') \left[h(\theta'|\theta) \min(1, \hat{\alpha}(\theta, \theta', y')) \right]. \end{aligned}$$

So we can write

$$\begin{aligned} (P - \hat{P})(\theta, d\theta') \\ = \delta_\theta(d\theta') \iint dt dy' h(t|\theta) F_t(y') \left[\min(1, \hat{\alpha}(\theta, t, y')) - \min(1, \alpha(\theta, t)) \right] \\ + \int dy' F_{\theta'}(y') \left[h(\theta'|\theta) \min(1, \alpha(\theta, \theta')) - h(\theta'|\theta) \min(1, \hat{\alpha}(\theta, \theta', y')) \right] \end{aligned}$$

and, finally,

$$\begin{aligned}
\|P - \hat{P}\| &= \frac{1}{2} \sup_{\theta} \int |P - \hat{P}|(\theta, d\theta') \\
&= \frac{1}{2} \sup_{\theta} \left\{ \left| \iint dt dy' h(t|\theta) F_t(y') [\min(1, \hat{\alpha}(\theta, t, y')) - \min(1, \alpha(\theta, t))] \right| \right. \\
&\quad \left. + \left| \iint dy' d\theta' F_{\theta'}(y') [h(\theta'|\theta) \min(1, \alpha(\theta, \theta')) - h(\theta'|\theta) \min(1, \hat{\alpha}(\theta, \theta', y'))] \right| \right\} \\
&= \sup_{\theta} \left\{ \left| \iint dt dy' h(t|\theta) F_t(y') [\min(1, \hat{\alpha}(\theta, t, y')) - \min(1, \alpha(\theta, t))] \right| \right\} \\
&\leq \sup_{\theta} \iint dy' d\theta' F_{\theta'}(y') h(\theta'|\theta) |\min(1, \alpha(\theta, \theta')) - \min(1, \hat{\alpha}(\theta, \theta', y'))| \\
&= \sup_{\theta} \int d\theta' h(\theta'|\theta) \int dy' F_{\theta'}(y') |\min(1, \alpha(\theta, \theta')) - \min(1, \hat{\alpha}(\theta, \theta', y'))| \\
&\leq \sup_{\theta} \int d\theta' h(\theta'|\theta) \delta(\theta, \theta'). \quad \square
\end{aligned}$$

Proof of Lemma 2.3. We still use Theorem 2.1, note that

$$\begin{aligned}
\|P_{\Sigma} - \hat{P}_{\Sigma}\| &= \frac{1}{2} \sup_{\theta} \int \left| \frac{1}{\sqrt{2\pi|\Sigma|}} \exp \left[-\frac{\|\Sigma^{-\frac{1}{2}}(\theta' - \theta - \frac{\Sigma}{2} \nabla \log \pi(\theta))\|^2}{2} \right] \right. \\
&\quad \left. - \frac{1}{\sqrt{2\pi|\Sigma|}} \exp \left[-\frac{\|\Sigma^{-\frac{1}{2}}(\theta' - \theta - \frac{\Sigma}{2} \hat{\nabla}^{y'} \log \pi(\theta))\|^2}{2} \right] \right| d\theta' F_{\theta}(dy') \\
&= \frac{1}{2} \sup_{\theta} \iint \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{\|t\|^2}{2} \right] \left| 1 \right. \\
&\quad \left. - \exp \left[\frac{\|t\|^2}{2} - \frac{\|t + \frac{\Sigma^{\frac{1}{2}}}{2} (\nabla \log \pi(\theta) - \hat{\nabla}^{y'} \log \pi(\theta))\|^2}{2} \right] \right| dt F_{\theta}(dy') \\
&= \frac{1}{2} \sup_{\theta} \iint \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{\|t\|^2}{2} \right] \left| 1 - \exp \left[\frac{t^T \Sigma^{\frac{1}{2}} (\nabla \log \pi(\theta) - \hat{\nabla}^{y'} \log \pi(\theta))}{2} \right] \right. \\
&\quad \left. - \frac{1}{8} \|\Sigma^{\frac{1}{2}} (\nabla \log \pi(\theta) - \hat{\nabla}^{y'} \log \pi(\theta))\|^2 \right| dt F_{\theta}(dy').
\end{aligned}$$

Now, note that

$$\begin{aligned}
& \int \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{\|t\|^2}{2} \right] \left| 1 \right. \\
& \quad \left. - \exp \left[\frac{t^T \Sigma^{\frac{1}{2}} (\nabla \log \pi(\theta) - \hat{\nabla}^{y'} \log \pi(\theta))}{2} - \frac{1}{8} \|\Sigma^{\frac{1}{2}} (\nabla \log \pi(\theta) - \hat{\nabla}^{y'} \log \pi(\theta))\|^2 \right] \right| dt \\
& \quad = \mathbb{E} \left| 1 - \exp \left(a^T X - \frac{\|a\|^2}{2} \right) \right|
\end{aligned}$$

where $X \sim \mathcal{N}(0, I)$ and $a = \Sigma^{\frac{1}{2}} [\nabla \log \pi(\theta) - \hat{\nabla}^{y'} \log \pi(\theta)]/2$. Then:

$$\begin{aligned}
\mathbb{E} \left| 1 - \exp \left(a^T X - \frac{\|a\|^2}{2} \right) \right| &= \exp \left(-\frac{\|a\|^2}{2} \right) \mathbb{E} \left| \exp(a^T X) - \exp \left(\frac{\|a\|^2}{2} \right) \right| \\
&= \exp \left(-\frac{\|a\|^2}{2} \right) \mathbb{E} \left| \exp(a^T X) - \mathbb{E} [\exp(a^T X)] \right| \\
&\leq \exp \left(-\frac{\|a\|^2}{2} \right) \sqrt{\text{Var}[\exp(a^T X)]} \\
&= \exp \left(-\frac{\|a\|^2}{2} \right) \sqrt{\mathbb{E} [\exp(2a^T X)] - \mathbb{E} [\exp(a^T X)]^2} \\
&= \exp \left(-\frac{\|a\|^2}{2} \right) \sqrt{\exp(2\|a\|^2) - \exp(\|a\|^2)} \\
&= \sqrt{\exp(\|a\|^2) - 1}.
\end{aligned}$$

So finally,

$$\begin{aligned}
\|P_\Sigma - \hat{P}_\Sigma\| &\leq \frac{1}{2} \sup_\theta \int F_\theta(dy') \sqrt{\exp \left[\frac{\|\Sigma^{\frac{1}{2}} (\nabla \log \pi(\theta) - \hat{\nabla}^{y'} \log \pi(\theta))\|^2}{4} \right] - 1} \\
&\leq \frac{1}{2} \sqrt{\sup_\theta \int F_\theta(dy') \exp \left[\frac{\|\Sigma^{\frac{1}{2}} (\nabla \log \pi(\theta) - \hat{\nabla}^{y'} \log \pi(\theta))\|^2}{4} \right] - 1} \leq \sqrt{\delta}. \quad \square
\end{aligned}$$

Proof of Lemma 3.1. We only have to check that

$$\begin{aligned}
& \mathbb{E}_{y' \sim F_{\theta'}} |\hat{\alpha}(\theta, \theta', y') - \alpha(\theta, \theta')| \\
& \leq \int dy' f(y'|\theta') \left| \alpha(\theta, \theta') - \hat{\alpha}(\theta, \theta', y') \right| \\
& = \frac{h(\theta|\theta')\pi(\theta')q_{\theta'}(y)}{h(\theta'|\theta)\pi(\theta)q_{\theta}(y)} \mathbb{E}_{y'_1, \dots, y'_N \sim f(\cdot|\theta')} \left| \frac{1}{N} \sum_{i=1}^N \frac{q_{\theta}(y'_i)}{q_{\theta'}(y'_i)} - \frac{Z(\theta)}{Z(\theta')} \right| \\
& \leq \frac{1}{\sqrt{N}} \frac{h(\theta|\theta')\pi(\theta')q_{\theta'}(y)}{h(\theta'|\theta)\pi(\theta)q_{\theta}(y)} \sqrt{\text{Var}_{y'_1 \sim f(y'_1|\theta')} \left(\frac{q_{\theta_n}(y'_1)}{q_{\theta'}(y'_1)} \right)}. \quad \square
\end{aligned}$$

Proof of Theorem 3.2. Under the assumptions of Theorem 3.2, note that (4) leads to

$$\alpha(\theta_n, \theta') = \frac{\pi(\theta')q_{\theta'}(y)Z(\theta_n)}{\pi(\theta_n)q_{\theta_n}(y)Z(\theta')} \frac{h(\theta_n|\theta')}{h(\theta'|\theta_n)} \geq \frac{1}{c_{\pi}^2 c_h^2 \mathcal{K}^4}. \quad (10)$$

Let us consider any measurable subset B of Θ and $\theta \in \Theta$. We have

$$\begin{aligned}
P(\theta, B) &= \int_B \delta_{\theta}(d\theta') \left[1 - \int dt h(t|\theta) \min(1, \alpha(\theta, t)) \right] \\
&\quad + \int_B d\theta' h(\theta'|\theta) \min(1, \alpha(\theta, \theta')) \\
&\geq \int_B d\theta' h(\theta'|\theta) \min(1, \alpha(\theta, \theta')) \\
&\geq \frac{1}{c_{\pi}^2 c_h^2 \mathcal{K}^4} \int_B d\theta' h(\theta'|\theta) \text{ thanks to (10)} \\
&\geq \frac{1}{c_{\pi}^2 c_h^3 \mathcal{K}^4} \int_B d\theta'.
\end{aligned}$$

This proves that Θ is a small set for the Lebesgue measure (multiplied by constant $1/c_{\pi}^2 c_h^3 \mathcal{K}^4$) on Θ . According to Theorem 16.0.2 page 394 in Meyn and Tweedie (Meyn and Tweedie 1993), this proves that:

$$\sup_{\theta} \|\delta_{\theta} P - \pi(\cdot|y)\| \leq C \rho^n$$

where

$$C = 2 \text{ and } \rho = 1 - \frac{1}{c_{\pi}^3 c_h^3 \mathcal{K}^4}$$

(note that, by definition, $\mathcal{K}, c_{\pi}, c_h > 1$ so we necessarily have $0 < \rho < 1$). So, Condition **(H1)** in Lemma 2.2 is satisfied.

Moreover,

$$\begin{aligned}\delta(\theta, \theta') &= \frac{h(\theta|\theta')\pi(\theta')q_{\theta'}(y)}{h(\theta'|\theta)\pi(\theta)q_{\theta}(y)} \sqrt{\text{Var}_{y' \sim f(y'|\theta')} \left(\frac{q_{\theta_n}(y')}{q_{\theta'}(y')} \right)} \\ &\leq c_h^2 c_\pi^2 \frac{q_{\theta'}(y)}{q_{\theta}(y)} \sqrt{\mathbb{E}_{y' \sim f(y'|\theta')} \left[\left(\frac{q_{\theta_n}(y')}{q_{\theta'}(y')} \right)^2 \right]} \leq c_h^2 c_\pi^2 \mathcal{K}^4.\end{aligned}$$

So, Condition **(H2)** in Lemma 2.2 is satisfied. We can apply this lemma and to give

$$\sup_{\theta_0 \in \Theta} \|\delta_{\theta_0} P^n - \delta_{\theta_0} \hat{P}^n\| \leq \frac{\mathcal{C}}{\sqrt{N}}$$

with

$$\mathcal{C} = c_\pi^2 c_h^2 \mathcal{K}^4 \left(\lambda + \frac{C\rho^\lambda}{1-\rho} \right)$$

with $\lambda = \left\lceil \frac{\log(1/C)}{\log(\rho)} \right\rceil$. \square

Proof of Lemma 3.3. Note that

$$\nabla \log \pi(\theta) - \hat{\nabla}^{x'} = \frac{1}{N} \sum_{i=1}^N s(y'_i) - \mathbb{E}_{y' \sim f_\theta}[s(y')].$$

So we have to find an upper bound, uniformly over θ , for

$$D := \mathbb{E}_{y' \sim F_{\theta_n}} \left\{ \exp \left[\frac{\sigma^2}{2} \left\| \Sigma^{\frac{1}{2}} \left(\frac{1}{N} \sum_{i=1}^N s(y'_i) - \mathbb{E}_{y' \sim f_\theta}[s(y')] \right) \right\|^2 \right] - 1 \right\}.$$

Let us put $V := \frac{1}{N} \sum_{i=1}^N V^{(i)} := \frac{1}{N} \sum_{i=1}^N \Sigma^{\frac{1}{2}} \{s(y'_i) - \mathbb{E}_{y' \sim f_\theta}[s(y')]\}$ and denote V_j ($j = 1, \dots, k$) the coordinates of V , and $V_j^{(i)}$ ($j = 1, \dots, k$) the coordinates of $V^{(i)}$. We have

$$\begin{aligned}D &= \mathbb{E} \left\{ \exp \left[\frac{1}{2} \sum_{j=1}^k V_j^2 \right] - 1 \right\} \\ &= \mathbb{E} \left\{ \exp \left[\frac{1}{k} \sum_{j=1}^k \frac{k}{2} V_j^2 \right] - 1 \right\} \\ &\leq \frac{1}{k} \sum_{j=1}^k \mathbb{E} \left\{ \exp \left[\frac{k}{2} V_j^2 \right] - 1 \right\}.\end{aligned}$$

Now, remark that $V_j = \frac{1}{N} \sum_{i=1}^n V_j^{(i)}$ with $-\mathcal{S}\|\Sigma\| \leq V_j^i \leq \mathcal{S}\|\Sigma\|$ so, Hoeffding's inequality ensures, for any $t \geq 0$,

$$\mathbb{P} \left(\left| \sqrt{N} V_j \right| \geq t \right) \leq 2 \exp \left[-\frac{t^2}{2\mathcal{S}^2 \|\Sigma\|^2} \right]$$

As a consequence, for any $\tau > 0$,

$$\begin{aligned}
\mathbb{E} \exp \left[\frac{k}{2} V_j^2 \right] &= \mathbb{E} \exp \left[\frac{k}{2N} \left(\sqrt{N} V_j \right)^2 \right] \\
&= \mathbb{E} \exp \left[\frac{k}{2N} \left(\sqrt{N} V_j \right)^2 \mathbf{1}_{|\sqrt{N} V_j| \leq \tau} \right] \\
&\quad + \mathbb{E} \exp \left[\frac{k}{2N} \left(\sqrt{N} V_j \right)^2 \mathbf{1}_{|\sqrt{N} V_j| > \tau} \right] \\
&= \exp \left(\frac{k\tau^2}{2N} \right) + \int_{\tau}^{\infty} \exp \left(\frac{k}{2N} x^2 \right) \mathbb{P} \left(|\sqrt{N} V_j| \geq x \right) dx \\
&\leq \exp \left(\frac{k\tau^2}{2N} \right) + 2 \int_{\tau}^{\infty} \exp \left[\left(\frac{k}{2N} - \frac{1}{2\mathcal{S}^2 \|\Sigma\|^2} \right) x^2 \right] dx \\
&= \exp \left(\frac{k\tau^2}{2N} \right) + 2 \sqrt{\frac{2\pi}{\frac{1}{\mathcal{S}^2 \|\Sigma\|^2} - \frac{2k}{N}}} \mathbb{P} \left(|\mathcal{N}| > \tau \sqrt{\frac{1}{\frac{1}{\mathcal{S}^2 \|\Sigma\|^2} - \frac{2k\sigma^2}{N}}} \right) \\
&\leq \exp \left(\frac{k\tau^2}{2N} \right) + 2 \sqrt{\frac{2\pi}{\frac{1}{\mathcal{S}^2 \|\Sigma\|^2} - \frac{2k}{N}}} \exp \left[-\frac{\tau^2}{\left(\frac{2}{\mathcal{S}^2 \|\Sigma\|^2} - \frac{4k}{N} \right)} \right] \\
&\leq \exp \left(\frac{k\tau^2}{2N} \right) + 2 \sqrt{\frac{2\pi}{\frac{1}{\mathcal{S}^2 \|\Sigma\|^2} - \frac{2k}{N}}} \exp \left[-\frac{\tau^2 \mathcal{S}^2 \|\Sigma\|^2}{2} \right]
\end{aligned}$$

where $\mathcal{N} \sim \mathcal{N}(0, 1)$. Now, we assume that $N > 4k\mathcal{S}^2 \|\Sigma\|^2$. This leads to $\frac{1}{\mathcal{S}^2 \|\Sigma\|^2} - \frac{2k}{N} > \frac{1}{2\mathcal{S}^2 \|\Sigma\|^2}$. This simplifies the bound to

$$\mathbb{E} \exp \left[\frac{k}{2} V_j^2 \right] \leq \exp \left(\frac{k\tau^2}{2N} \right) + 4\sqrt{\pi} \mathcal{S} \|\Sigma\| \exp \left[-\frac{\tau^2 \mathcal{S}^2 \|\Sigma\|^2}{2} \right].$$

Finally, we put $\tau = \sqrt{\log(N/k)/(2\mathcal{S}^2 \|\Sigma\|^2)}$ to get

$$\mathbb{E} \exp \left[\frac{k}{2} V_j^2 \right] \leq \exp \left(\frac{k \log \left(\frac{N}{k} \right)}{4\mathcal{S}^2 \|\Sigma\|^2 N} \right) + \frac{4k\sqrt{\pi} \mathcal{S} \|\Sigma\|}{N}.$$

It follows that

$$D \leq \exp \left(\frac{k \log(N)}{4\mathcal{S}^2 \|\Sigma\|^2 N} \right) - 1 + \frac{4k\sqrt{\pi} \mathcal{S} \|\Sigma\|}{N}.$$

This ends the proof. \square